Published in final edited form as:
Genet Epidemiol. 2013 September ; 37(6): 560–570. doi:10.1002/gepi.21740.

Haplotype Kernel Association Test as a Powerful Method to Identify Chromosomal Regions Harboring Uncommon Causal Variants

Wan-Yu Lin[1,*], **Nengjun Yi**[2], **Xiang-Yang Lou**[2], **Degui Zhi**[2], **Kui Zhang**[2], **Guimin Gao**[3], **Hemant K. Tiwari**[2], and **Nianjun Liu**[2,*]

[1]Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan

[2]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama

[3]Department of Biostatistics, Virginia Commonwealth University, Richmond, Virginia

Abstract

For most complex diseases, the fraction of heritability that can be explained by the variants discovered from genome-wide association studies is minor. Although the so-called 'rare variants' (minor allele frequency [MAF] < 1%) have attracted increasing attention, they are unlikely to account for much of the 'missing heritability' because very few people may carry these rare variants. The genetic variants that are likely to fill in the 'missing heritability' include uncommon causal variants (MAF < 5%), which are generally untyped in association studies using tagging single-nucleotide polymorphisms (SNPs) or commercial SNP arrays. Developing powerful statistical methods can help to identify chromosomal regions harboring uncommon causal variants, while bypassing the genome-wide or exome-wide next-generation sequencing. In this work, we propose a haplotype kernel association test (*HKAT*) that is equivalent to testing the variance component of random effects for distinct haplotypes. With an appropriate weighting scheme given to haplotypes, we can further enhance the ability of *HKAT* to detect uncommon causal variants. With scenarios simulated according to the population genetics theory, *HKAT* is shown to be a powerful method for detecting chromosomal regions harboring uncommon causal variants.

Keywords

Similarity; Linkage disequilibrium; Rare variants; *JAK2* gene; Body-mass index

[*]Corresponding authors: Nianjun Liu, Ph.D., Ryals Public Health Bldg 327, 1665 University Blvd, University of Alabama at Birmingham, Birmingham, AL 35294-0022, U.S.A., Phone: (205) 975-9190, Fax: (205) 975-2540, nliu@uab.edu; Wan-Yu Lin, Ph.D., Room 501, No. 17, Xu-Zhou Road, Taipei 100, Taiwan, Phone: +886-2-33668106, Fax: +886-2-33668106, linwy@ntu.edu.tw.

The authors declare that they have no conflict of interest.

Introduction

Genetic association studies have provided insights into the genetic architecture of complex diseases [WTCCC, 2007; Hardy and Singleton, 2009]. However, for most complex diseases, the fraction of heritability that can be explained by the variants discovered from association studies remains minor [Maher, 2008; Manolio et al., 2009; Eichler et al., 2010; Gibson, 2012]. Although the so-called ‘rare variants’ (minor allele frequency [MAF] < 1%) have attracted increasing attention, they are unlikely to account for much of the ‘missing heritability’ because very few people may carry these rare variants [Pihur and Chakravarti, 2010]. The best bet of genetic variants to fill in the ‘missing heritability’ includes two sources: uncommon causal variants (MAF < 5%) that are generally untyped in association studies using tagging single-nucleotide polymorphisms (SNPs) or commercial SNP arrays, and common causal variants with small genetic effects that cannot be detected via conventional statistical analyses [Manolio et al., 2009; Eichler et al., 2010; Yi et al., 2011]. Indeed, existing association studies such as genome-wide association studies (GWAS) or candidate-gene association studies (CGAS) are not designed to capture uncommon causal variants [Wray et al., 2011]. The emergence of next-generation sequencing technologies has allowed for the mapping of all genetic variants across the human genome [Hawkins et al., 2010]. However, the cost of sequencing remains high [Sboner et al., 2011]. Genome-wide sequencing is especially expensive for large sample sizes that are required for association studies [Sampson et al., 2012]. In the current stage, GWAS and CGAS data are still much more widely available than next-generation sequencing data [WTCCC, 2007; Li et al., 2010].

The widely used single-marker analysis that is implemented on each tagging SNP (usually with MAF ≥ 5%) is underpowered for detecting uncommon causal variants [Gusev et al., 2011], because the information of uncommon causal variants is not easy to be represented by common SNPs. Haplotypes, combinations of multiple adjacent alleles on a single chromosome, may act as ‘superalleles’ and serve as better tagging markers for uncommon causal variants that are generally not genotyped in GWAS or CGAS [Lin et al., 2012b]. For case-control studies with unrelated subjects, haplotype frequencies are often compared between cases and controls with a likelihood-ratio statistic [Zhao et al., 2000; Epstein and Satten, 2003; Becker et al., 2005]. To deal with continuous traits, a regression framework has been introduced to relate inferred haplotype frequencies to observed phenotypes [Zaykin et al., 2002]. Moreover, score tests based on generalized linear models have been proposed to deal with a variety of traits [Schaid et al., 2002]. Methods with use of haplotype similarity [Tzeng et al., 2003] and haplotype clustering [Molitor et al., 2003; Durrant et al., 2004; Tzeng, 2005; Tzeng et al., 2006; Browning and Browning, 2007] were also developed for GWAS or CGAS.

Modelling individual effects for all distinct haplotypes may induce many parameters and cause computation problems to the conventional likelihood-ratio test [Schaid et al., 2002]. In this work, we propose a haplotype kernel association test (*HKAT*) that is equivalent to testing the variance component of random effects for distinct haplotypes. Despite a large number of distinct haplotypes in a region, the signal of haplotype-trait association can be aggregated to a single variance parameter. With an appropriate weighting scheme given to haplotypes, we can further enhance the ability of *HKAT* to detect uncommon causal variants. We also consider the situation that a gene or a chromosomal region harbors not only uncommon causal variants but also common causal variants, and then we compare *HKAT* with several popular genotype or haplotype analysis methods by performing systematic simulations under a wide range of linkage disequilibrium (LD) patterns. In addition, we apply *HKAT* to data from a genetic association study related to human adiposity.

Materials and Methods

Haplotype Kernel Association Test (HKAT)

Let Y_i be the trait of the i^{th} subject (i = 1, …, n), and let $\boldsymbol{x_i}$=[$x_{i,1}$ $x_{i,2}$ $\cdots$ $x_{i,p}$]' be a vector that codes p non-genetic covariates (e.g., age, gender, ethnicity, etc.) of the i^{th} subject. To account for haplotype ambiguity, the expectation-maximization algorithm [Dempster et al., 1977] is often used to infer the posterior distribution of haplotypes given multimarker genotypes. Let $\boldsymbol{h_i}$=[$h_{i,1}$ $h_{i,2}$ $\cdots$ $h_{i,L}$]' be the i^{th} subject's expected frequencies of L distinct haplotypes over his/her posterior distribution of haplotypes. To relate the genetic composition to the trait, we consider a linear model for a continuous trait:

$$E(Y_i)=\alpha_0+\boldsymbol{\alpha}'\boldsymbol{x}_i+\boldsymbol{\beta}'\boldsymbol{h}_i, \quad (1)$$

or a logistic regression model for a dichotomous trait:

$$logit\ P(Y_i=1)=\alpha_0+\boldsymbol{\alpha}'\boldsymbol{x}_i+\boldsymbol{\beta}'\boldsymbol{h}_i, \quad (2)$$

where α_0 is the intercept term, $\boldsymbol{\alpha}$ = [α_1 α_2 $\cdots$ α_p]' is the vector of regression coefficients for the p covariates, and $\boldsymbol{\beta}$ = [β_1 β_2 $\cdots$ β_L]' is the vector of regression coefficients for the L distinct haplotypes.

To test if any of the haplotypes are associated with the trait, the null hypothesis is H_0 : $\boldsymbol{\beta}$ = **0**, i.e., H_0 : β_1 = β_2 =$\cdots$ = β_L = 0. However, the commonly used likelihood-ratio test is computationally intensive and underpowered especially when some haplotypes are of low frequency. To reduce the number of parameters for distinct haplotypes, we assume that β_j is a random effect following an arbitrary distribution with a mean of zero and a variance of $w_j\tau$, where τ is a variance component and w_j is a pre-specified weight for the jth distinct haplotype. Therefore, τ is a common parameter for all of the distinct haplotypes and w_j's (j = 1, …, L) are pre-specified weights for these distinct haplotypes. To test whether the regression coefficients of the L distinct haplotypes are all zero (H_0 : $\boldsymbol{\beta}$ = **0**) is equivalent to test whether the variance component is zero (H_0 : τ = 0). The score statistic to test H_0 : τ = 0 is

$$T_{HKAT}=(\boldsymbol{y}-\hat{\mu})'\boldsymbol{H}'\boldsymbol{W}_H\boldsymbol{H}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}), \quad (3)$$

where $\boldsymbol{y}$ is the vector of traits of all the n subjects, $\hat{\boldsymbol{\mu}}$ is the predicted mean of $\boldsymbol{y}$ under the null hypothesis (H_0 : τ = 0), $\boldsymbol{H}$ is the haplotype frequency matrix with the i^{th} column as $\boldsymbol{h_i}$, and

$\boldsymbol{W_H}$ is a diagonal matrix with the (j, j)-th element to be the pre-specified weight for the jth distinct haplotype (w_j). This test is referred to as the haplotype kernel association test (*HKAT*).

According to the theory of quadratic forms of normal variables [Scheffe, 1959], T_{HKAT} is asymptotically distributed as a mixture of χ^2 variables: $\sum_{i=1}^{\varpi} \lambda_i \chi^2_{1,i}$, where $\chi^2_{1,i}$'s are independent χ^2 variables with one degree of freedom, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\varpi}$ are the ordered eigenvalues of the matrix $\boldsymbol{P}_0^{1/2} \boldsymbol{H}' \boldsymbol{W}_{\boldsymbol{H}} \boldsymbol{H} \boldsymbol{P}_0^{1/2}$ (with the rank of ϖ). To reduce the bias that may be caused by a small sample size, we use the restricted maximum likelihood estimator of the variance component [Zhang and Lin, 2003] and therefore the matrix $\boldsymbol{P}_0 = \hat{\boldsymbol{V}}^{-1} - \hat{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}} (\tilde{\boldsymbol{X}}' \hat{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}' \hat{\boldsymbol{V}}^{-1}$, where $\tilde{\boldsymbol{X}} = [\boldsymbol{1}\ \boldsymbol{X}]$ is a $n \times (p + 1)$ matrix and $\hat{\boldsymbol{V}}$ is a diagonal matrix with the (i, i)-th element to be the estimated variance of $\hat{\mu}_i$. For a continuous trait, $\hat{\boldsymbol{V}} = \hat{\sigma}_0^2 \boldsymbol{I}$ where $\hat{\sigma}_0^2$ is the mean squared error under the null hypothesis and $\boldsymbol{I}$ is an $n \times n$ identity matrix. For a dichotomous trait, $\hat{\boldsymbol{V}} = \boldsymbol{diag}\ (\hat{\mu}_1(1 - \hat{\mu}_1), \hat{\mu}_2(1 - \hat{\mu}_2), \cdots, \hat{\mu}_n(1 - \hat{\mu}_n))$ where $\hat{\mu}_i = logit^{-1}\ (\hat{\alpha}_0 + \hat{\alpha}' \boldsymbol{x_i})$ is the estimated probability of being a case under the null hypothesis. The distribution of T_{HKAT} can be approximated by the three-moment approximation method [Imhof, 1961; Zhang, 2005; Allen and Satten, 2007, 2009; Pan, 2009; Tzeng et al., 2009], and the P-value of the observed *HKAT* test statistic is given by

$$P\left(\chi_b^2 > (T_{HKAT} - c_1) \times \sqrt{\frac{b}{c_2}} + b\right), \quad (4)$$

where $c_j = \sum_{i=1}^{\varpi} \lambda_i^j$, $b = c_2^3 \big/ c_3^2$, and χ_b^2 is the χ^2 distribution with b degrees of freedom.

Genotype Kernel Association Test (GKAT)

To investigate the association of genetic variants in a chromosomal region with the disease, we can use genotypes to bypass the haplotype-phasing stage. Let $\boldsymbol{g_i}$ be a vector of genotype scores of the i^{th} subject at the set of markers in the chromosomal region. Under the assumption of additive genetic model, the possible elements of $\boldsymbol{g_i}$ are 0, 1, and 2, representing the number of copies of the minor allele. The vector $\boldsymbol{g_i}$ can be recoded accordingly if dominant or recessive genetic models are considered. In Equations (1) and (2), if we substitute $\boldsymbol{h_i}$ with $\boldsymbol{g_i}$, the score statistic to test whether the variance component of genotypes is zero will be

$$T_{GKAT} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})' \boldsymbol{G}' \boldsymbol{W}_G \boldsymbol{G} (\boldsymbol{y} - \hat{\boldsymbol{\mu}}), \quad (5)$$

where $\boldsymbol{G}$ is the genotype matrix with the i^{th} column to be $\boldsymbol{g_i}$, and $\boldsymbol{W_G}$ is a diagonal matrix with the (j, j)-th element to be the weight given to the jth genetic variant. This test is referred to as the genotype kernel association test (*GKAT*). Similarly, T_{HKAT} is asymptotically distributed as a mixture of χ^2 variables: $\sum_{i=1}^{\varpi'} \lambda_i \chi^2_{1,i}$, where $\chi^2_{1,i}$'s are independent χ^2

variables with one degree of freedom, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\varpi'}$ are the ordered eigenvalues of the matrix $\boldsymbol{P}_0^{1/2}\boldsymbol{G}'\boldsymbol{W}_G\boldsymbol{G}\boldsymbol{P}_0^{1/2}$ (with the rank of ϖ').

The test statistic of *GKAT* is equivalent to that of the popular sequence kernel association test (referred to as '*SKAT*') [Wu et al., 2011], except the weight given to genetic variants, i.e. $\boldsymbol{W_G}$ in Equation (5). In *SKAT*, the weight given to the *j*th variant is $w_j = Beta\ (p_j;\ a_{1,}\ a_2)^2$, where p_j is the MAF of the *j*th variant, and $a_{1,}$ and a_2 are suggested to be set at 1 and 25, respectively [Wu et al., 2011]. We call the test in Equation (5) *GKAT* rather than *SKAT* [Wu et al., 2011], because we want to distinguish the situations of using *SKAT* and *GKAT*. *SKAT* has been proposed by Wu et al. [2011] for analyzing sequencing data, whereas *GKAT* is used to analyze genotyped SNPs in GWAS or CGAS.

$\boldsymbol{W_H}$ and $\boldsymbol{W_G}$

In Equations (3) and (5), $\boldsymbol{W_H}$ and $\boldsymbol{W_G}$ are diagonal matrices with weights given to distinct haplotypes and SNPs, respectively. If mutations are rare, the distribution of the frequency (p) of the mutant allele is $f(p) \propto p^{-1}$ [Wright, 1931; Crow and Kimura, 1970; Kimura, 1983; Hill et al., 2008]. A causal allele may be a mutant allele or an ancestral allele, so the frequencies of causal alleles follow a *U*-shaped distribution; i.e. $f(p) \propto p^{-1} + (1-p)^{-1} = [p(1-p)]^{-1}$ [Hill et al., 2008]. Therefore, a straightforward weight given to a genetic variant with MAF of p_j is $[p_j\,(1-p_j)]^{-1}$. To avoid obtaining an extreme weight given a p_j very close to 0, we follow Madsen and Browning [2009] to estimate frequencies as $\hat{p}_j = \frac{(m_j+1)}{(2n_j+2)}$, where m_j is the number of minor allele observed for the *j*th SNP and n_j is the total number of subjects genotyped for that SNP. In the following, *GKAT* is evaluated with $\boldsymbol{W_G} = diag\ ([\hat{p}_1(1-\hat{p}_1)]^{-\mathrm{k}}, \cdots, [\hat{p}_L(1-\hat{p}_L)]^{-\mathrm{k}})$, where L is the number of loci in the chromosomal region and $k=0, \frac{1}{2}$, and 1, respectively. According to the different levels of *k,* the test is referred to as *GKAT0*, *GKAT1/2*, or *GKAT1*, respectively. As mentioned above, $k = 1$ is a straightforward choice given the *U*-shaped distribution for a causal allele [Hill et al., 2008].

The choice of $k=\frac{1}{2}$ is based on Madsen and Browning's [2009] weight given to genetic variants. In addition, $k = 0$ represents a same weight given to all variants, regardless of their MAFs.

Parallel to *GKAT*, *HKAT* is evaluated at $\boldsymbol{W_H} = diag([\hat{f}_1(1-\hat{f}_1)]^{-k}, \cdots, [\hat{f}_L(1-\hat{f}_L)]^{-k})$, where L here is the number of distinct haplotypes in the chromosomal region and $k=0, \frac{1}{2}$, and 1, respectively. The test is referred to as *HKAT0*, *HKAT1/2*, or *HKAT1*, respectively. In $\boldsymbol{W_H}$, f_h is the frequency of haplotype h, estimated with $\hat{f}_h = \frac{(C_h+1)}{(2n+2)}$, where C_h is the number of haplotype h among all of the n subjects. When haplotype phases are ambiguous, C_h can be inferred from unphased multimarker genotypes using the expectation-maximization algorithm [Dempster et al., 1977], under the assumption of Hardy-Weinberg equilibrium [Excoffier and Slatkin, 1995; Hawley and Kidd, 1995; Long et al., 1995].

When dealing with case-control studies, some researchers [Madsen and Browning, 2009; Li et al., 2010; Lin et al., 2012b] have proposed using only unaffected subjects to estimate MAFs or haplotype frequencies. However, weights dependent on traits (affected or unaffected) will inflate type-I error rates [Lin and Tang, 2011], especially for the *HKAT1* test. Suppose the count of some distinct haplotype is five in the pooled sample and that, *by chance*, these five haplotypes are all contributed by the affected subjects. If we calculate the frequencies and consequent weights with only the unaffected subjects, this haplotype will be even more up-weighted (i.e., more than if it were weighted independently of the traits) and this *artificial* association will be amplified. This phenomenon will jeopardize the validity of the *HKAT1* test, in which a larger magnitude of weight (k = 1) is given to haplotypes. Therefore, we use the whole sample to estimate MAFs (in *GKAT*) or haplotype frequencies (in *HKAT*).

Simulation study

Following Li et al.'s simulation [2010] and using the Cosi program [Schaffner et al., 2005], we generated 500 data sets each containing 10,000 chromosomes of 1 Mb regions. The chromosomes were generated according to the LD patterns of the HapMap CEU samples, and an ~50 kb causal region was randomly picked from the 1 Mb region for each data set. Within each causal region, we randomly selected d variants (d = 5, 10, 20, 30, or 40) as causal variants. When evaluating the performance of different methods for detecting uncommon causal variants, the causal variants were chosen from the variants with population MAFs ranging from 0.1% to 5%. In addition, a gene may harbor both uncommon and common causal variants, and therefore we also consider the scenario with causal variants having population MAFs ranging from 0.1% to 30%. Minor alleles were treated as causal alleles, which might be deleterious or protective (or, increase or decrease the trait values, when continuous traits are simulated). We let r_{isk} % of the d causal variants increase the disease risk, while the remaining (1- r_{isk}) % decrease the disease risk (or increase/decrease the value of a continuous trait). The value of r_{isk} was evaluated at 5, 20, 50, 80, and 100, respectively. To mimic the selection of tagging SNPs based on the HapMap CEU data, for each data set, we randomly chose 120 from the 10,000 chromosomes and paired them as 60 subjects. Based on the LD patterns of the 60 subjects, we used the *H-clust* method [Rinaldo et al., 2005; Roeder et al., 2005] to select tagging SNPs with the conventional criteria, i.e. $r^2 > 0.8$ (only one SNP selected from a group of SNPs in LD with $r^2 > 0.8$) and MAF > 5% [Barrett and Cardon, 2006; Keating et al., 2008]. These tagging SNPs were served as genotyped genetic variants in our simulations. For each simulated data set, a 20-tagging-SNP window that encompasses the causal region was chosen as a multimarker set used for analysis.

Dichotomous traits

Population genetics theories and empirical studies all support the assumption that the effect sizes of causal variants tend to be inversely related to their allele frequencies [Park et al. 2011; Bodmer and Bonilla, 2008; Eyre-Walker, 2010; Weetman et al., 2010; Ramsey et al., 2012]. Therefore, following previous studies [Madsen and Browning, 2009; Li et al., 2010; Lin et al., 2012b], we let the genotype relative risk (GRR) of the jth causal variant be

$$GRR_j=\left(\frac{PAR_j}{(1-PAR_j)\cdot MAF_j}+1\right)^{(-1)^{I(\xi_j=1)}}, \quad (6)$$

where PAR_j and MAF_j are the population attributable risk (PAR) and the population MAF of the jth causal variant, respectively. The indicator function $I(\xi_j = 1)$ is 1 or 0 according to whether the jth causal variant is protective or deleterious. Given PAR, the relationship between MAF and GRR is shown in Supplementary Figure S1. In addition, Supplementary Figures S2 and S3 present the distributions of MAFs and GRRs of the causal variants in our 500 simulated data sets, respectively.

To generate the genotypes of an individual, we randomly selected two chromosomes from the remaining 9,880 (= 10,000–120) chromosomes. The disease status of an individual with chromosomes $\{H_1, H_2\}$ was determined by

$$P(affected|\{H_1,H_2\})=f_0\times\prod_{k=1}^{2}\prod_{j=1}^{d}GRR_j^{I(H_{k,j}=a_j)}, \quad (7)$$

where f_0 was the baseline penetrance and was fixed at 10% [Li et al., 2010; Lin et al., 2012b], and a_j was the minor allele of the jth causal variant. The total sample size was set at 2,000. Considering that cases are usually more difficult to recruit and so many studies have fewer cases than controls [WTCCC, 2007; Zhernakova et al., 2007; Barrett et al., 2011; Macgregor et al., 2011; Sawcer et al., 2011], we let the 2,000 subjects be composed of 400 cases and 1,600 controls (a balanced case-control design with equal numbers of cases and controls will be discussed later). After generating the disease status based on Equation (7), the genotypes of the causal variants that were not selected as tagging SNPs were removed from our analysis data sets. When all of the causal variants were uncommon (MAF < 5%), almost all of them were removed from the multimarker set because the tagging SNPs were selected with the criterion of MAF > 5% [Barrett and Cardon, 2006; Keating et al., 2008]. When the causal variants were selected from those having MAFs ∈[0.1%, 30%], some common causal variants (MAF > 5%) might be reserved in the multimarker set if they were selected as tagging SNPs.

Continuous traits

In addition to dichotomous traits, we also simulated continuous traits. The trait value (Y) was generated by

$$Y=10C_1+10C_2+\beta_1 g_1+\beta_2 g_2+\cdots+\beta_d g_d+e, \quad (8)$$

where C_1 was a continuous covariate following a standard normal distribution, C_2 was a dichotomous covariate taking a value of 0 or 1 each with a probability of 0.5, g_j was the number of causal allele on the j^{th} causal variant (g_j = 0, 1, or 2), β_j was the effect size of the j^{th} causal variant, and e was the random error. The random error, e, was assumed to have a normal distribution with a mean of zero and a variance of V_e. The effect sizes βs and V_e

were determined so that the ‘marginal heritability’ (the heritability of each causal variant, notated as h^2 and $h^2 = \frac{Var(\beta_j g_j)}{Var(Y)} = \frac{Var(\beta_j g_j)}{Var(10C_1 + 10C_2) + dVar(\beta_j g_j) + V_e}$ for $j = 1,\dots,d$) was fixed at 0.05%, 0.1%, 0.15%, or 0.2% under the alternative hypothesis. The actual values of V_e and βs were not critical. Once V_e was specified, βs were determined via the setting of the marginal heritability. We first assigned an arbitrary value to V_e, and we then obtained $\beta_j (j = 1,\dots,d)$ from

$$Var(\beta_j g_j) = \beta_j^2 \cdot 2 \cdot MAF_j \cdot (1 - MAF_j) = \frac{h^2 \cdot [\, V_e + Var(10C_1 + 10C_2)]}{1 - d \cdot h^2} = \frac{h^2 \cdot (V_e + 125)}{1 - d \cdot h^2}. \quad (9)$$

The relationship between βs and the MAFs of causal variants is shown in Supplementary Figure S4. The total sample size was set at 2,000. After generating the traits, the genotypes of the causal variants that were not selected as tagging SNPs were removed from our analysis data sets.

Tests under comparison

We compared the three *HKAT* tests (*HKAT0*, *HKAT1/2*, *HKAT1*) and the three *GKAT* tests (*GKAT0*, *GKAT1/2*, *GKAT1*) with a global score test for haplotypes (hereinafter referred to as ‘*global*’) and a test based on the maximum score statistic over all haplotypes (hereinafter referred to as ‘*max*’), both of which have been widely used for haplotype association analyses [Schaid et al., 2002]. The *global* tests the overall effect of all haplotypes, while *max* tests the effect of the most significant haplotype. When performing *global* and *max*, the haplotypes with counts less than 5 were lumped into a single baseline group, according to the default of the package ‘haplo.stats’ [Schaid et al., 2002]. To allow the *HKAT* tests to be robust to genotyping errors, we merged haplotypes having a count less than 5 with their most similar haplotypes having a count larger than 5, where ‘5’ was chosen to lead to a parallel comparison on *HKAT*, *global*, and *max*. Under the assumption of Hardy-Weinberg equilibrium [Excoffier and Slatkin, 1995; Hawley and Kidd, 1995; Long et al., 1995], we used the ‘haplo.em’ function in the ‘haplo.stats’ package [Schaid et al., 2002] to infer haplotype phases from unphased multimarker genotypes with the expectation-maximization algorithm [Dempster et al., 1977]. The *j*th element of $\boldsymbol{h_i} = [h_{i,1}\ h_{i,2} \cdots h_{i,L}]'$ in Equation (1) is determined by $h_{i,j} = \frac{1}{2}\sum_{k \neq j} \Pr(H_j, H_k | \boldsymbol{g_i}) + \Pr(H_j, H_j | \boldsymbol{g_i})$, where $\Pr(H_j, H_k \mid \boldsymbol{g_i})$ is the posterior distribution of haplotype pairs (H_j, H_k)given multimarker genotypes $\boldsymbol{g_i}$. In this way, all possible haplotype pairs were considered with their posterior probabilities. To have a better control of type-I error rates, phasing cases and controls together (instead of phasing them separately) was suggested [Lin and Huang, 2007]. Therefore, we phased the pooled sample of cases and controls when dichotomous traits were evaluated.

In addition to *global* and *max*, we used the R package ‘SKAT’ to perform the popular sequence kernel association test (referred to as ‘*SKAT*’) [Wu et al., 2011], as well as the optimal test (referred to as ‘*SKAT-Op*’) [Lee et al., 2012], which optimally combines the burden tests [Li and Leal, 2008; Madsen and Browning, 2009; Morris and Zeggini, 2010; Price et al., 2010; Lin et al., 2011] and *SKAT* [Wu et al., 2011]. Both *SKAT* and *SKAT-Op* were proposed for dealing with sequencing data, therefore we applied these two approaches to the full sequence (rather than merely the 20 tagging SNPs) of the analysis region. For any given data set, there were around 170∼280 observed variants in an analysis region. With consideration of cost, there is a trade-off between the number of subjects and different study designs (CGAS or next generation sequencing) [Sampson et al., 2011; Sboner et al., 2011]. Therefore, following a suggestion from an anonymous reviewer, when performing *SKAT* and *SKAT-Op* on full sequencing data, the total sample size was set at 200 (or 40 cases and 160 controls for simulations of dichotomous traits) rather than 2000.

When analyzing dichotomous traits, we also included a haplotype grouping test (referred to as '*HG*') [Feng and Zhu, 2010; Zhu et al., 2010] and a weighted haplotype test on genotyped SNPs (referred to as '*WHG*') [Li et al., 2010] into comparisons. First, the data are split into a training set and a testing set. *HG* classifies haplotypes as risk or non-risk with the training set, and then tests for associations by performing a Fisher's exact test with the testing set. *WHG* is based on a similar procedure, but it further boosts power to detect rare variants by weighting haplotypes according to their frequencies. For both tests, we randomly selected 30% of the sample as the training set and let the remaining 70% be the testing set, following the allocation chosen by previous studies [Li et al., 2010; Lin et al., 2012b].

Results

Type-I error rates

By setting the PAR (for dichotomous traits) or the marginal heritability (for continuous traits) at exactly 0%, we evaluated type-I error rates by performing 1,000 replications for each of the 500 simulated data sets. The *P*-values of *global* and *max* were obtained with 1,000-20,000 permutations by a sequential Monte Carlo algorithm [Besag and Clifford, 1991], according to the default of the package 'haplo.stats' [Schaid et al., 2002]. Then we evaluated type-I error rates given significance levels from 10^{-4} to 10^{-1}. Based on 500,000 (=500×1000) replications across the 500 simulated data sets, Figure 1 shows that all of the 12 tests (for dichotomous traits) or 10 tests (for continuous traits) are valid in the sense that their type-I error rates match the nominal significance levels.

Power comparisons

When we evaluated power, a total of 100 replications were performed under each scenario (each combination of r_{isk}, PAR or marginal heritability, and *d*) for each of the 500 simulated data sets. Figures 2 and 3 present the power averaged over the 500 data sets, given a nominal significance level of 10^{-3}, for dichotomous traits and continuous traits, respectively. When the nominal significance level is set at 10^{-4}, we get the results presented in Supplementary Figures S5-S6. *HKAT1* (*HKAT* with a weighting order $k = 1$) is the most powerful test, given uncommon causal variants with MAFs ∈[0.1%, 5%] or given a mixture of uncommon and common causal variants with MAFs ∈[0.1%, 30%].

The power performance of these tests may be sensitive to (1) the percentage of rare variants among all causal variants, and (2) the LD pattern between the causal variants and the surrounding markers. With stratified analysis, we find that *HKAT1* consistently outperforms other tests over all ranges of percentage of rare variants, and all ranges of average r^2 between causal variants and surrounding markers (data not shown).

Regarding the power performance of different levels of weighting order, $k = 1$ is the best, followed by $k=\frac{1}{2}$ and $k = 0$, for both *HKAT* and *GKAT*. This is because $k = 1$ setting up-weights rare haplotypes that are more likely to tag rare causal variants. As can be seen in the top rows of Figures 2-3, genotype-based tests (*GKAT* and *SKAT* [Wu et al., 2011; Lee et al., 2012], which are equivalent except for different weighting schemes given to variants) are underpowered when all causal variants are uncommon with population MAFs $\in$[0.1%, 5%], because their power can only be driven by tagging SNPs (usually with MAF > 5% [Barrett and Cardon, 2006; Keating et al., 2008]) that are generally not good surrogates for uncommon causal variants. Haplotype-based tests (*HKAT*, *global*, and *max*) are more powerful because haplotypes can be better tags for uncommon causal variants. When some causal variants are common (so that the tagging SNPs are likely to represent the information of these common causal variants), the performance of genotype-based tests (*GKAT* and *SKAT*) improves, although it still cannot compete with *HKAT* (see the bottom rows of Figures 2-3). Note that our results for *SKAT* and *SKAT-Op* do not imply that they have similar performance when dealing with next-generation sequencing data in which rare variants should also be genotyped. Here we incorporated these two tests simply because of their relatedness with *GKAT*.

Application to a Human Adiposity Study

Next, we applied the 10 tests for continuous traits to a human adiposity study [Chung et al., 2009]. In this study, 1,982 unrelated European Americans living in the New York City metropolitan area were recruited. We investigated the association of 17 tagging SNPs in the *Janus kinase 2* (*JAK2*) gene (located on chromosome 9p24) with body-mass index (BMI). These 17 tagging SNPs were selected from SNPs from 10,000 base pairs upstream to 10,000 base pairs downstream of *JAK2*'s coding sequence, according to the conventional criteria of $r^2 > 0.8$ and MAF > 5%. Following Chung et al. [2009], we first adjusted the log-transformed BMI with sex, age, age^2, and their respective interactions. Associations of the joint additive and dominance effects of each of the 17 tagging SNPs with BMI were tested using the ordinary-least-squares regression method. Consistent with the results from Chung et al. [2009] (see their Table 3), there were six SNPs with *P*-values smaller than 0.05, with the smallest *P*-value (0.008) being observed on SNP *rs3780365*. However, after correcting for multiple testing, none of the six SNPs was significant at the family-wise error rate of 0.05.

We then resorted to the 10 multimarker tests. The first step was to define a ‘multimarker set’. A natural strategy is to aggregate all SNPs located in a gene [Schifano et al., 2012]. We let all the 17 SNPs in the *JAK2* gene be a ‘multimarker set’ and analyze this set with the 10 multimarker tests, respectively. Among the 10 tests, *GKAT0*, *GKAT1/2*, *GKAT1*, and *HKAT1* suggest that the *JAK2* gene is associated with BMI, and the *P*-values are 0.025, 0.026, 0.024, and 0.025, respectively. The *P*-values of other six multimarker tests are all larger than 0.05.

Another commonly used strategy is to partition a gene into segments according to the LD patterns [Gabriel et al., 2002; Zhang et al., 2002; Twells et al., 2003; Han and Pan, 2010; Schifano et al., 2012]. Based on the default of Haploview [Barrett et al., 2005] to customize the haplotype blocks (the Gabriel et al.'s rule [2002]), there are two haplotype blocks in the *JAK2* gene. We applied the 10 multimarker tests to the two haplotype blocks, respectively. Only *HKAT1* and *global* suggest an association of haplotypes from the second block (*rs3780365- rs2230724- rs1410779- rs3824432- rs3780372- rs10491652- rs3780379- rs966871*) with BMI, and the *P*-values are 0.004 and 0.006, respectively.

JAK2 is involved in leptin, insulin, and ABCA1 (the adenosine triphosphate-binding cassette transporter A1) signaling pathways [Banks et al., 2000]. Disturbance in leptin and insulin signalling pathways are related to obesity and metabolic syndrome [Penas-Steinhardt et al., 2011]. It may influence body fat mass, insulin sensitivity, or serum lipid profile in humans [Ge et al., 2008]. An independent study genotyped tagging SNPs spanning *JAK2* for 2,760 white female twin subjects from the St. Thomas' U.K. Adult Twin Registry [Ge et al., 2008], and it led to a similar result as that of Chung et al.'s study [2009]. That is, although some *P*-values of SNP-obesity association were smaller than 0.05, none of these remained significant after adjusting for multiple testing. Investigation of the tagging SNPs via *HKAT* may provide additional information that may be missed by single-marker analyses.

Discussion

Because the cost of sequencing remains high, association studies using SNP arrays or tagging SNPs are still among the most commonly available data types in the current stage [WTCCC, 2007; Li et al., 2010]. The aim of this study is to provide a valid and powerful statistical method for detecting disease-associated genomic regions with uncommon causal variants from contemporary GWAS or CGAS data sets, thus bypassing the genome-wide or exome-wide next-generation sequencing. Both uncommon causal variants with large effect sizes and common variants with small effect sizes are possible to contribute to the missing heritability for complex diseases [Manolio et al., 2009; Eichler et al., 2010; Yi et al., 2011]. Single-locus analysis is underpowered to detect these two types of causal variants [Stahl et al., 2010, 2012]. Because a susceptibility gene is likely to harbor multiple causal variants [Hugot et al., 2001; Ogura et al., 2001; Pritchard, 2001; WTCCC, 2007; Madsen and Browning, 2009; Wang et al., 2010], we investigate methods that can test multiple SNPs aggregately for a collective signal on traits. These methods include *SKAT*, which is popular and powerful for rare variant detection [Wu et al., 2010; Wu et al., 2011; Lee et al., 2012]; *global* and *max* [Schaid et al., 2002], which have been widely used for detecting haplotype-trait association; our *HKAT* and *GKAT* equipped with three levels of weighting order ($k=0, \frac{1}{2}$, and 1). After simulating scenarios based on the population genetics theory [Wright, 1931; Crow and Kimura, 1970; Kimura, 1983; Hill et al., 2008], we find that *HKAT1* is the best test to detect the signal of uncommon causal variants.

HKAT is computationally feasible, because it is based on a score test without fitting the full model (i.e., the model under the alternative hypothesis). On an Intel Xeon workstation with 3.0 GHz of CPU and 2.0 GB of memory, *HKAT* with a 20-SNP multimarker set on average takes ~0.9, ~7.0, and ~22.8 seconds to analyze 1000, 2000, and 3000 subjects, respectively. In genetic studies, haplotype phase is usually unknown when diploid subjects are heterozygous at more than one chromosomal locus. Therefore, we inferred haplotype information with the expectation-maximization algorithm [Dempster et al., 1977], which leads inferred haplotype frequencies to maximum likelihood estimates under the assumption of Hardy-Weinberg equilibrium [Excoffier and Slatkin, 1995; Hawley and Kidd, 1995; Long et al., 1995]. There are two common uses of the inferred haplotypes in downstream analyses. One way is to use the most likely haplotype pair, which has the highest posterior probability among all possible haplotype pairs of a subject. The most likely haplotype pair is assigned probability 1 and all other possible haplotype pairs are assigned probabilities 0. This common practice is intrinsically biased because the most likely haplotype pair is not necessarily the true haplotype pair of that subject [Lin and Huang, 2007]. Another way is the expectation substitution approach [Zaykin et al., 2002; Stram et al., 2003]. That is, a subject's expected frequencies of haplotypes are treated as observed and directly used in downstream analyses. Under the null hypothesis of no haplotype effects, similar to previous methods [Schaid et al., 2002; Zaykin et al., 2002; Stram et al., 2003], the resulting score statistic (i.e., HKAT statistic in Equation (3)) is unbiased and gives correct type-I error rates (see Figure 1). Employing this expectation substitution approach, although the variability of the estimated haplotype frequencies is not explicitly incorporated in the variance of the HKAT statistic, the HKAT test is shown to be valid.

The *HKAT* and *GKAT* proposed here are applicable to CGAS or GWAS. An issue is how to select a set of SNPs to be included in a multimarker test. Natural strategies include aggregating all SNPs located in a gene or within a haplotype block [Feng and Zhu, 2010; Lin et al., 2012a; Schifano et al., 2012], as we have shown in the analysis for the human adiposity study. Haplotype-based methods such as *HKAT* are justifiable to analyze haplotype blocks, which are discrete chromosome regions containing SNPs in high LD [Cardon and Abecasis, 2003]. Another strategy is to use sliding windows [Guo et al., 2009; Wang et al., 2012]. In general, multimarker analyses with larger window sizes may allow for measuring sharing over longer genomic sequences and lead to more power gains [Allen and Satten, 2009; Lin et al., 2012b].

The *HKAT* and *GKAT* can be applicable to continuous or dichotomous traits. In our simulation for dichotomous traits, we considered an unbalanced case-control design with 20% cases and 80% controls. For a balanced case-control design (with 50% cases and 50% controls), *HKAT1* has a similar performance with *global*. However, *HKAT1* is still more advantageous than *global* in computational feasibility, because no permutation is required for *HKAT1*. By contrast, *global* needs permutations to obtain reliable *P*-values when the frequencies of some haplotypes are low [Schaid et al., 2002; Lin et al., 2012b].

Our work shows that in GWAS using commercial SNP arrays or CGAS using tagging SNPs, the haplotype-based methods (e.g., *HKAT*, *global* [Schaid et al., 2002], *max* [Schaid et al., 2002], *HG* [Feng and Zhu, 2010; Zhu et al., 2010], and *WHG* [Li et al., 2010]) are more promising than the genotype-based methods (e.g., *GKAT*, *SKAT* [Wu et al., 2011], and *SKAT-Op* [Lee et al., 2012]) in detecting uncommon causal variants. Among haplotype-based methods, *HKAT* is further shown to be more powerful than *HG* [Feng and Zhu, 2010;

Zhu et al., 2010] and *WHG* [Li et al., 2010], because the power of *HG* or *WHG* is generally compromised due to splitting the data into two subsets (i.e., a training set and a testing set). In addition, *HKAT1* outperforms *global* and *max* by up-weighting uncommon haplotypes that may be better tags for uncommon causal variants. When a gene harbors both uncommon and common causal variants, *HKAT1* remains the most powerful test among all the tests we evaluate here. Note that this conclusion is based on the simulation scenario following the population genetics theory (i.e., the distribution of causal allele frequencies is *U*-shaped) [Wright, 1931; Crow and Kimura, 1970; Kimura, 1983; Hill et al., 2008], and in this situation $k = 1$ is a straightforward and reasonable weighting order. However, for any given study, the most powerful test may vary if the underlying genetic architecture departs from the population genetics theory.

At the pseudo-sequencing level (i.e., GWAS or CGAS imputed with publicly available sequencing data) [Li et al., 2010] or the sequencing level, the haplotype-based methods may not be as promising as the genotype-based methods. This deserves further investigation. In recent years, many novel methods have been proposed for rare variant identification using next-generation sequencing data [Li and Leal, 2008; Madsen and Browning, 2009; Han and Pan, 2010; Liu and Leal, 2010; Morris and Zeggini, 2010; Price et al., 2010; Basu and Pan, 2011; Lin et al., 2011; Neale et al., 2011; Wu et al., 2011; Yi et al., 2011; Yi and Zhi, 2011; Lee et al., 2012; Liu and Leal, 2012]. However, next-generation sequencing data have not been prevalent till today. By contrast, few methods have been proposed for detecting uncommon causal variants from genetic association studies genotyped with tagging SNPs or commercial SNP arrays. We here provide a haplotype-based test that is powerful to detect disease-associated regions from GWAS or CGAS.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the anonymous reviewers for their insightful and constructive comments and Drs. Wendy K. Chung and Rudolph L. Leibel for kindly providing the adiposity data (sponsored by NIH grant DK52431 and the New York Health Project). This work was supported by NSC grant 102-2314-B-002-001-MY2 from the National Science Council of Taiwan (W-Y.L.) and NIH grants GM081488 (N.L.), 5R01GM069430-08 (N.Y.), DA025095 (X.L.), R00 RR024163 (D.Z.), and GM073766 (G.G.) from the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors.

References

Allen AS, Satten GA. Statistical models for haplotype sharing in case-parent trio data. Hum Hered. 2007; 64(1):35–44. [PubMed: 17483595]

Allen AS, Satten GA. A novel haplotype-sharing approach for genome-wide case-control association studies implicates the calpastatin gene in Parkinson's disease. Genet Epidemiol. 2009; 33(8):657–67. [PubMed: 19365859]

Banks AS, Davis SM, Bates SH, Myers MG Jr. Activation of downstream signals by the long form of the leptin receptor. J Biol Chem. 2000; 275(19):14563–72. [PubMed: 10799542]

Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. Nat Genet. 2006; 38(6):659–62. [PubMed: 16715099]

Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005; 21(2):263–5. [PubMed: 15297300]

Barrett JH, Iles MM, Harland M, Taylor JC, Aitken JF, Andresen PA, Akslen LA, Armstrong BK, Avril MF, Azizi E, et al. Genome-wide association study identifies three new melanoma susceptibility loci. Nat Genet. 2011; 43(11):1108–13. [PubMed: 21983787]

Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. Genet Epidemiol. 2011; 35(7):606–19. [PubMed: 21769936]

Becker T, Cichon S, Jonson E, Knapp M. Multiple testing in the context of haplotype analysis revisited: application to case-control data. Ann Hum Genet. 2005; 69(6):747–56. [PubMed: 16266412]

Besag J, Clifford P. Sequential Monte Carlo p-values. Biometrika. 1991; 78:301–304.

Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008; 40(6):695–701. [PubMed: 18509313]

Browning BL, Browning SR. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. Genet Epidemiol. 2007; 31(5):365–375. [PubMed: 17326099]

Cardon LR, Abecasis GR. Using haplotype blocks to map human complex trait loci. Trends Genet. 2003; 19(3):135–40. [PubMed: 12615007]

Chung WK, Patki A, Matsuoka N, Boyer BB, Liu N, Musani SK, Goropashnaya AV, Tan PL, Katsanis N, Johnson SB, et al. Analysis of 30 genes (355 SNPS) related to energy homeostasis for association with adiposity in European-American and Yup'ik Eskimo populations. Hum Hered. 2009; 67(3):193–205. [PubMed: 19077438]

Crow, JF.; Kimura, M. An Introduction to Population Genetics Theory. New York: Harper & Row; 1970.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood estimation from incomplete data via the EM algorithm. J R Stat Soc. 1977; 39:1–38.

Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet. 2004; 75(1):35–43. [PubMed: 15148658]

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010; 11(6):446–50. [PubMed: 20479774]

Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet. 2003; 73(6):1316–29. [PubMed: 14631556]

Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol. 1995; 12(5):921–7. [PubMed: 7476138]

Eyre-Walker A. Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc Natl Acad Sci U S A. 2010; 107(Suppl 1):1752–6. [PubMed: 20133822]

Feng T, Zhu X. Genome-wide searching of rare genetic variants in WTCCC data. Hum Genet. 2010; 128(3):269–80. [PubMed: 20549515]

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, et al. The structure of haplotype blocks in the human genome. Science. 2002; 296(5576):2225–9. [PubMed: 12029063]

Ge D, Gooljar SB, Kyriakou T, Collins LJ, Swaminathan R, Snieder H, Spector TD, O'Dell SD. Association of common JAK2 variants with body fat, insulin sensitivity and lipid profile. Obesity (Silver Spring). 2008; 16(2):492–6. [PubMed: 18239666]

Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2012; 13(2):135–45. [PubMed: 22251874]

Guo Y, Li J, Bonham AJ, Wang Y, Deng H. Gains in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: a comparison of association-mapping strategies. Eur J Hum Genet. 2009; 17(6):785–92. [PubMed: 19092774]

Gusev A, Kenny EE, Lowe JK, Salit J, Saxena R, Kathiresan S, Altshuler DM, Friedman JM, Breslow JL, Pe'er I. DASH: A Method for Identical-by-Descent Haplotype Mapping Uncovers Association with Recent Variation. Am J Hum Genet. 2011; 88(6):706–17. [PubMed: 21620352]

Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered. 2010; 70(1):42–54. [PubMed: 20413981]

Hardy J, Singleton A. Genomewide association studies and human disease. N Engl J Med. 2009; 360(17):1759–68. [PubMed: 19369657]

Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet. 2010; 11(7):476–86. [PubMed: 20531367]

Hawley ME, Kidd KK. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered. 1995; 86(5):409–11. [PubMed: 7560877]

Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. 2008; 4(2):e1000008. [PubMed: 18454194]

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature. 2001; 411(6837):599–603. [PubMed: 11385576]

Imhof JP. Computing the distribution of quadratic forms in normal variables. Biometrika. 1961; 48:419–426.

Keating BJ, Tischfield S, Murray SS, Bhangale T, Price TS, Glessner JT, Galver L, Barrett JC, Grant SF, Farlow DN, et al. Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. PLoS One. 2008; 3(10):e3583. [PubMed: 18974833]

Kimura, M. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press; 1983.

Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012; 13(4):762–75. [PubMed: 22699862]

Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83(3):311–21. [PubMed: 18691683]

Li Y, Byrnes AE, Li M. To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests. Am J Hum Genet. 2010; 87(5):728–35. [PubMed: 21055717]

Lin DY, Huang BE. The use of inferred haplotypes in downstream analyses. Am J Hum Genet. 2007; 80(3):577–9. [PubMed: 17380613]

Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet. 2011; 89(3):354–67. [PubMed: 21885029]

Lin WY, Tiwari HK, Gao G, Zhang K, Arcaroli JJ, Abraham E, Liu N. Similarity-based multimarker association tests for continuous traits. Ann Hum Genet. 2012a; 76(3):246–60. [PubMed: 22497480]

Lin WY, Yi N, Zhi D, Zhang K, Gao G, Tiwari HK, Liu N. Haplotype-based methods for detecting uncommon causal variants with common SNPs. Genet Epidemiol. 2012b; 36(6):572–82. [PubMed: 22706849]

Lin WY, Zhang B, Yi N, Gao G, Liu N. Evaluation of pooled association tests for rare variant identification. BMC Proc. 2011; 5(Suppl 9):S118. [PubMed: 22373333]

Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet. 2010; 6(10):e1001156. [PubMed: 20976247]

Liu DJ, Leal SM. A unified framework for detecting rare variant quantitative trait associations in pedigree and unrelated individuals via sequence data. Hum Hered. 2012; 73(2):105–22. [PubMed: 22555759]

Long JC, Williams RC, Urbanek M. An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet. 1995; 56(3):799–810. [PubMed: 7887436]

Macgregor S, Montgomery GW, Liu JZ, Zhao ZZ, Henders AK, Stark M, Schmid H, Holland EA, Duffy DL, Zhang M, et al. Genome-wide association study identifies a new melanoma susceptibility locus at 1q21.3. Nat Genet. 2011; 43(11):1114–8. [PubMed: 21983785]

Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5(2):e1000384. [PubMed: 19214210]

Maher B. Personal genomes: The case of the missing heritability. Nature. 2008; 456(7218):18–21. [PubMed: 18987709]

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747–53. [PubMed: 19812666]

Molitor J, Marjoram P, Thomas D. Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. Am J Hum Genet. 2003; 73(6):1368–84. [PubMed: 14631555]

Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010; 34(2):188–93. [PubMed: 19810025]

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. PLoS Genet. 2011; 7(3):e1001322. [PubMed: 21408211]

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature. 2001; 411(6837):603–6. [PubMed: 11385577]

Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol. 2009; 33(6):497–507. [PubMed: 19170135]

Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni JF Jr, Chatterjee N. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci U S A. 2011; 108(44):18026–31. [PubMed: 22003128]

Penas-Steinhardt A, Tellechea ML, Gomez-Rosso L, Brites F, Frechtel GD, Poskus E. Association of common variants in JAK2 gene with reduced risk of metabolic syndrome and related disorders. BMC Med Genet. 2011; 12:166. [PubMed: 22185674]

Pihur, V.; Chakravarti, A. Neither rare nor common variants can explain much of phenotypic variation. Presented at the 60th Annual Meeting of the American Society of Human Genetics; November 5, 2010; Washington, D.C.. 2010.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010; 86(6):832–8. [PubMed: 20471002]

Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet. 2001; 69(1):124–37. [PubMed: 11404818]

Ramsey LB, Bruun GH, Yang W, Trevino LR, Vattathil S, Scheet P, Cheng C, Rosner GL, Giacomini KM, Fan Y, et al. Rare versus common variants in pharmacogenetics: SLCO1B1 variation and methotrexate disposition. Genome Res. 2012; 22(1):1–8. [PubMed: 22147369]

Rinaldo A, Bacanu SA, Devlin B, Sonpar V, Wasserman L, Roeder K. Characterization of multilocus linkage disequilibrium. Genet Epidemiol. 2005; 28(3):193–206. [PubMed: 15637716]

Roeder K, Bacanu SA, Sonpar V, Zhang X, Devlin B. Analysis of single-locus tests to detect gene/disease associations. Genet Epidemiol. 2005; 28(3):207–19. [PubMed: 15637715]

Sampson J, Jacobs K, Yeager M, Chanock S, Chatterjee N. Efficient study design for next generation sequencing. Genet Epidemiol. 2011; 35:269–277.

Sampson JN, Jacobs K, Wang Z, Yeager M, Chanock S, Chatterjee N. A two-platform design for next generation genome-wide association studies. Genet Epidemiol. 2012; 36(4):400–8. [PubMed: 22508365]

Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, Dilthey A, Su Z, Freeman C, Hunt SE, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011; 476(7359):214–9. [PubMed: 21833088]

Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! Genome Biol. 2011; 12(8):125. [PubMed: 21867570]

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 2005; 15(11):1576–83. [PubMed: 16251467]

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet. 2002; 70(2):425–34. [PubMed: 11791212]

Scheffe, H. The Analysis of Variance. New York: Wiley; 1959.

Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SL, Peyser PA, Lin X. SNP Set Association Analysis for Familial Data. Genet Epidemiol. 2012; 36:797–810.

Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li Y, Kurreeman FA, Zhernakova A, Hinks A, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet. 2010; 42(6):508–14. [PubMed: 20453842]

Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, Kraft P, Chen R, Kallberg HJ, Kurreeman FA, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet. 2012; 44(5):483–9. [PubMed: 22446960]

Stram DO, Leigh Pearce C, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered. 2003; 55(4): 179–90. [PubMed: 14566096]

Twells RC, Mein CA, Phillips MS, Hess JF, Veijola R, Gilbey M, Bright M, Metzker M, Lie BA, Kingsnorth A, et al. Haplotype structure, LD blocks, and uneven recombination within the LRP5 gene. Genome Res. 2003; 13(5):845–55. [PubMed: 12727905]

Tzeng JY. Evolutionary-based grouping of haplotypes in association analysis. Genet Epidemiol. 2005; 28(3):220–31. [PubMed: 15726584]

Tzeng JY, Devlin B, Wasserman L, Roeder K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet. 2003; 72(4):891–902. [PubMed: 12610778]

Tzeng JY, Wang CH, Kao JT, Hsiao CK. Regression-based association analysis with clustered haplotypes through use of genotypes. Am J Hum Genet. 2006; 78(2):231–42. [PubMed: 16365833]

Tzeng JY, Zhang D, Chang SM, Thomas DC, Davidian M. Gene-trait similarity regression for multimarker-based association analysis. Biometrics. 2009; 65(3):822–32. [PubMed: 19210740]

Wang K, Dickson SP, Stolle CA, Krantz ID, Goldstein DB, Hakonarson H. Interpretation of association signals and identification of causal variants from genome-wide association studies. Am J Hum Genet. 2010; 86(5):730–42. [PubMed: 20434130]

Wang X, Liu X, Sim X, Xu H, Khor CC, Ong RT, Tay WT, Suo C, Poh WT, Ng DP, et al. A statistical method for region-based meta-analysis of genome-wide association studies in genetically diverse populations. Eur J Hum Genet. 2012; 20(4):469–75. [PubMed: 22126751]

Weetman D, Wilding CS, Steen K, Morgan JC, Simard F, Donnelly MJ. Association mapping of insecticide resistance in wild Anopheles gambiae populations: major variants identified in a low-linkage disequilbrium genome. PLoS One. 2010; 5(10):e13140. [PubMed: 20976111]

Wray NR, Purcell SM, Visscher PM. Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol. 2011; 9(1):e1000579. [PubMed: 21267061]

Wright S. Evolution in Mendelian Populations. Genetics. 1931; 16(2):97–159. [PubMed: 17246615]

WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447(7145):661–78. [PubMed: 17554300]

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010; 86(6):929–42. [PubMed: 20560208]

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89(1):82–93. [PubMed: 21737059]

Yi N, Liu N, Zhi D, Li J. Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects. PLoS Genet. 2011; 7(12):e1002382. [PubMed: 22144906]

Yi N, Zhi D. Bayesian analysis of rare variants in genetic association studies. Genet Epidemiol. 2011; 35(1):57–69. [PubMed: 21181897]

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered. 2002; 53(2):79–91. [PubMed: 12037407]

Zhang D, Lin X. Hypothesis testing in semiparametric additive mixed models. Biostatistics. 2003; 4(1):57–74. [PubMed: 12925330]

Zhang JT. Approximate and asymptotic distributions of Chi-squared-type mixtures with applications. Journal of the American Statistical Association. 2005; 100:273–285.

Zhang K, Calabrese P, Nordborg M, Sun F. Haplotype block structure and its applications to association studies: power and study designs. Am J Hum Genet. 2002; 71(6):1386–94. [PubMed: 12439824]

Zhao JH, Curtis D, Sham PC. Model-free analysis and permutation tests for allelic associations. Hum Hered. 2000; 50(2):133–9. [PubMed: 10799972]

Zhernakova A, Alizadeh BZ, Bevova M, van Leeuwen MA, Coenen MJ, Franke B, Franke L, Posthumus MD, van Heel DA, van der Steege G, et al. Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. Am J Hum Genet. 2007; 81(6):1284–8. [PubMed: 17999365]

Zhu X, Feng T, Li Y, Lu Q, Elston RC. Detecting rare variants for complex traits using family and unrelated data. Genet Epidemiol. 2010; 34(2):171–87. [PubMed: 19847924]

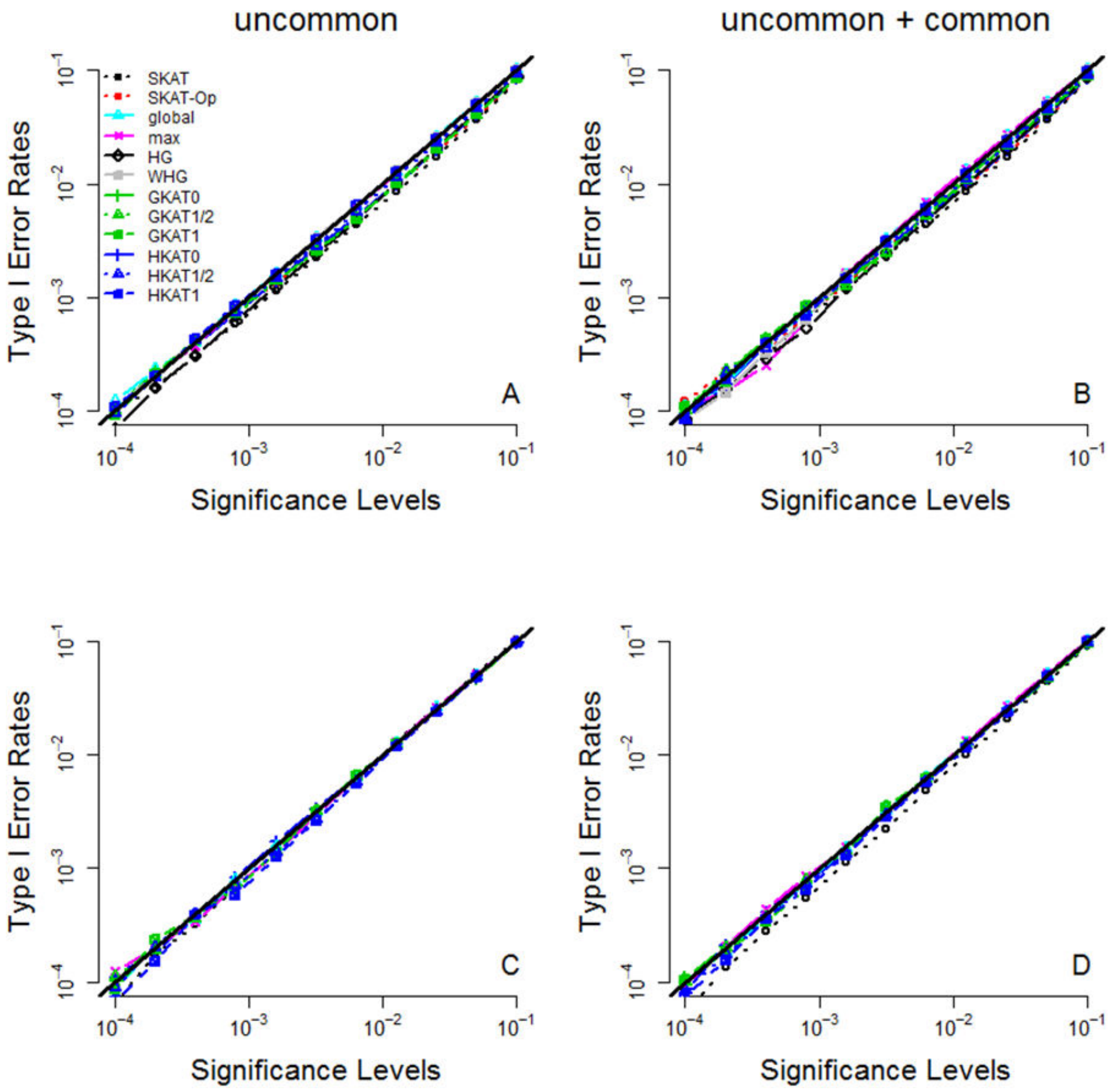

Figure 1. Type-I error rates

The *x*-axis is the nominal significance level (where the leftmost point is 10^{-4} and the rightmost point is 10^{-1}), and the *y*-axis is the type-I error rate. Panel (A): dichotomous traits and ‘uncommon’ causal variants with MAFs ∈[0.1%, 5%]; panel (B): dichotomous traits and ‘uncommon + common’ causal variants with MAFs ∈[0.1%, 30%]; panel (C): continuous traits and ‘uncommon’ causal variants with MAFs ∈[0.1%, 5%]; panel (D): continuous traits and ‘uncommon + common’ causal variants with MAFs ∈[0.1%, 30%]. The curves of all the tests are on the line $y = x$ (the black bold line).

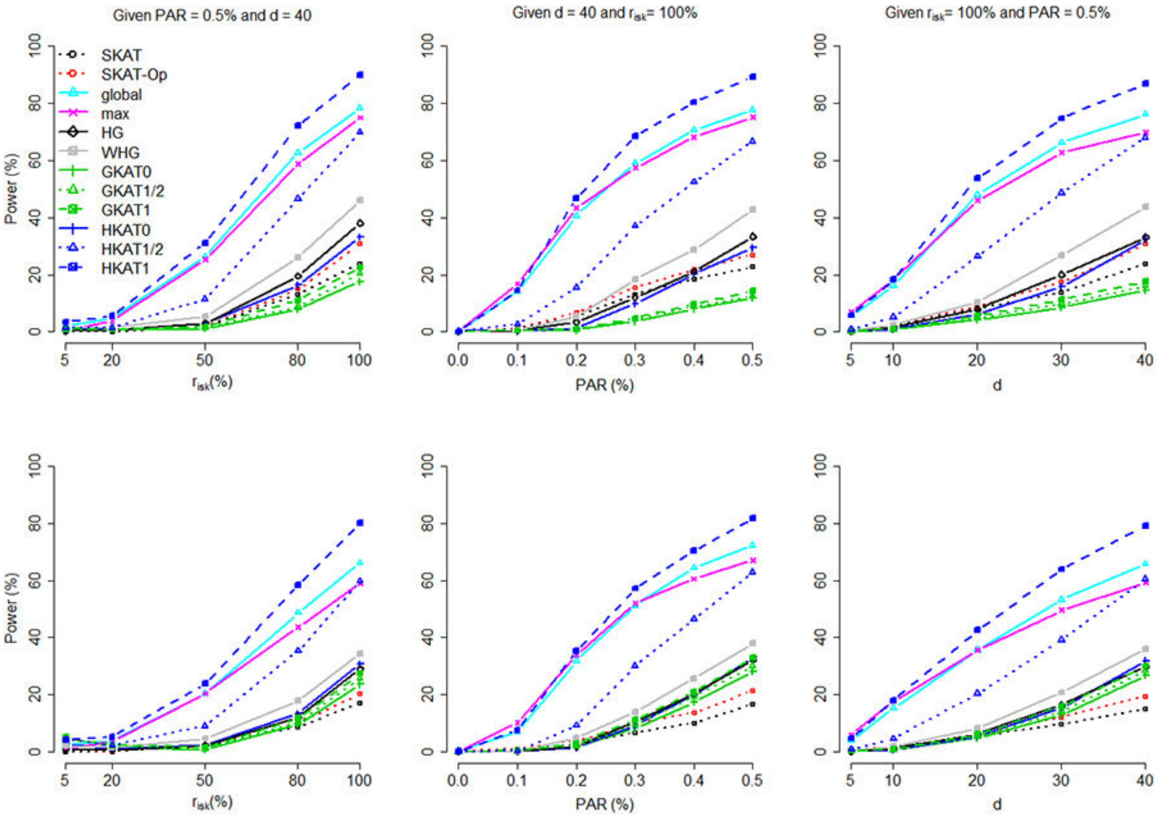

Figure 2. Dichotomous trait - Comparison of power by r_{isk} (the percentage of deleterious variants among the *d* causal variants), PAR, and *d* (the number of causal variants)

The figure shows the power comparison by r_{isk} (left column, given PAR = 0.5% and d = 40), PAR (middle column, given d = 40 and r_{isk} = 100%), and d (right column, given r_{isk} = 100% and PAR = 0.5%), respectively. The nominal significance level was set at 10^{-3}. The top row is the result given 'uncommon' causal variants with MAFs ∈[0.1%, 5%]; the bottom row is the result given 'uncommon + common' causal variants with MAFs ∈[0.1%, 30%].

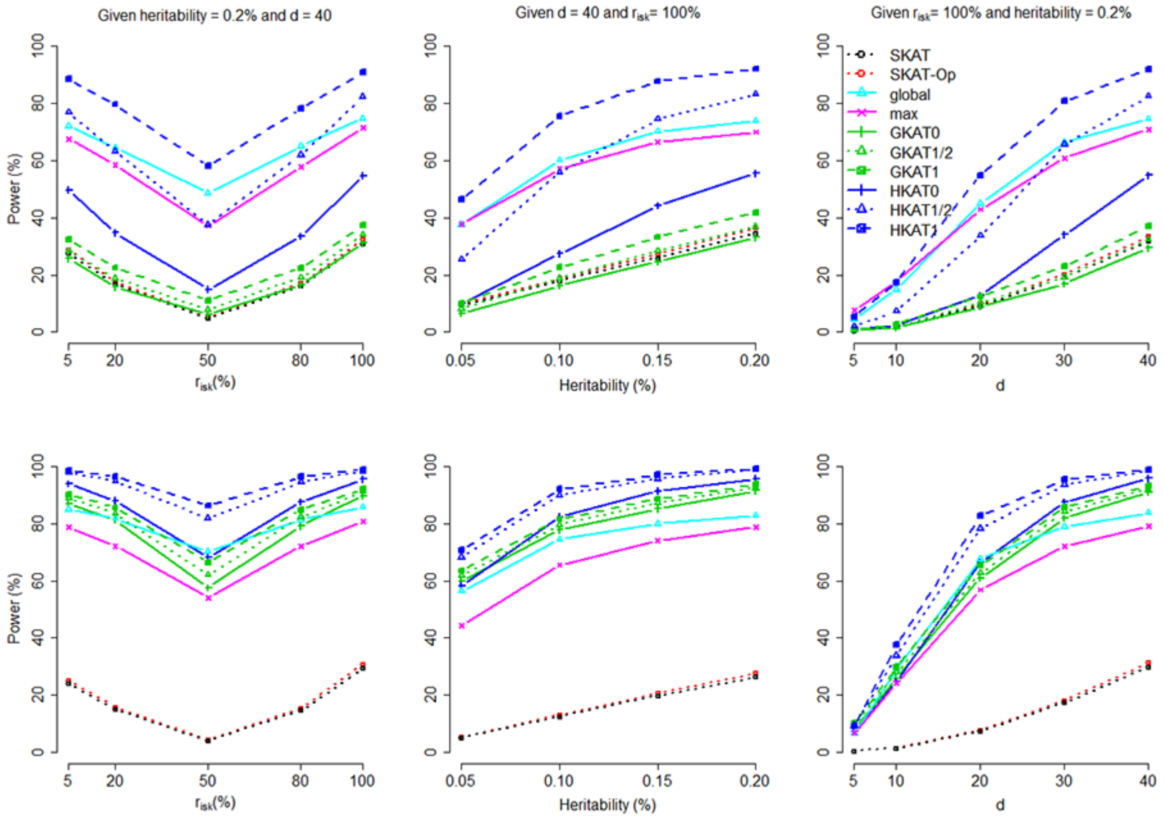

Figure 3. Continuous trait - Comparison of power by r_{isk} (the percentage of variants among the *d* causal variants that increase the trait value), the marginal heritability, and *d* (the number of causal variants)

The figure shows the power comparison by r_{isk} (left column, given the marginal heritability = 0.2% and d = 40), the marginal heritability (middle column, given d = 40 and r_{isk} = 100%), and d (right column, given r_{isk} = 100% and the marginal heritability = 0.2%), respectively. The nominal significance level was set at 10^{-3}. The top row is the result given 'uncommon' causal variants with MAFs ∈[0.1%, 5%]; the bottom row is the result given 'uncommon + common' causal variants with MAFs ∈[0.1%, 30%].